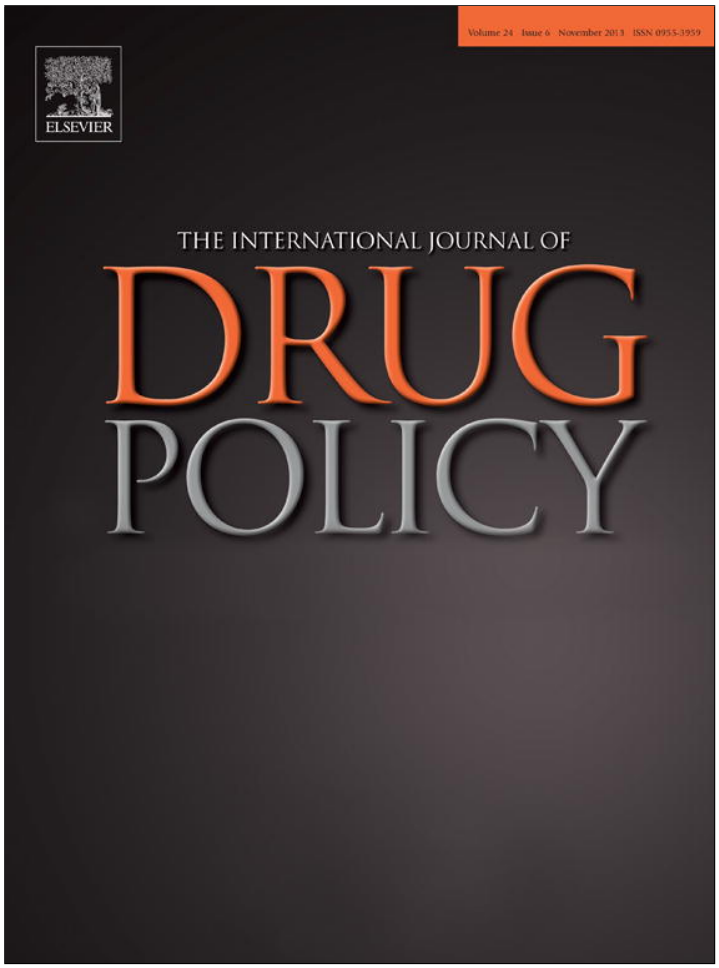

International Journal of Drug Policy 24 (2013) 614–618

Contents lists available at ScienceDirect

International Journal of Drug Policy

journal homepage: www.elsevier.com/locate/drugpo

Research paper

Methods for population size estimation of problem drug users using a single registration

Peter G.M. van der Heijden[a,*], Guus Cruts[b], Maarten Cruyff[a]

[a] Department of Methodology and Statistics, Faculty of Social and Behavioural Sciences, Utrecht University, PO Box 80140, NL-3508 TC Utrecht, The Netherlands
[b] Trimbos Institute, Netherlands Institute of Mental Health and Addiction, PO Box 725, NL-3500 AS Utrecht, The Netherlands

ARTICLE INFO

Article history:
Received 6 February 2012
Received in revised form 12 July 2012
Accepted 10 April 2013

Keywords:
Estimation
Single registration capture recapture
Problem drug users

ABSTRACT

Background: The number of problem drug users is used as a key indicator to monitor the drug situation in the European Union. An alternative approach to estimate the number of problem drug users is given by 'the one-source capture–recapture analysis' that uses a single registration.
Methods: Two variants of the one-source capture–recapture analysis were applied to a single registration: the truncated Poisson regression model (TPR) and the Zelterman regression model. These models were applied to data about clinical drug-related hospital admissions derived from the Dutch Hospital Registration (LMR). The TPR accounts for heterogeneity in capture probabilities by allowing for covariates and the Zelterman regression model relies on the problem drug users that were seen only once or twice in the hospital; the latter model is known to be robust against unobserved heterogeneity.
Results: The TPR model was found to have a bad fit due to unobserved heterogeneity. The Zelterman regression model estimated the population size at 10,415 problem drug users (95% CI is 8400–12,429). This figure is an estimate of the number of problem drug users who are at risk of a clinical hospital admission due to the medical consequences of their drug use. The model can also provide estimates of different subgroups of problematic drug users.
Conclusion: The method presented here offers a promising alternative for estimating the number of problem drug users, including different subgroups of drug users. In addition, observed and unobserved heterogeneity can be accounted for in these estimates.

Introduction

The European Monitoring Centre for Drugs and Drug Addiction (EMCDDA, 2011) uses the number of problem drug users as a key indicator to evaluate the effects of different drug policies in the European Union (EU). When classifying these drug policies, a distinction is usually made between repressive and more liberal policies that are based upon harm reduction. In order to answer the question about whether a repressive or more liberal drug policy is most effective at reducing the number of problem drug users, a valid estimate of their number in any jurisdiction is required.

In the Netherlands, a member state of the EU, estimates of the number of problem drug users have been obtained mainly by means of the treatment multiplier and the two-source capture–recapture analysis. Most recently, the treatment multiplier used a field sample and a treatment registration (Cruts & Van Laar, 2010), and the two-source capture–recapture analysis used police arrest data and a treatment registration sample (Temürhan et al., submitted for publication).

Here we propose the use of one-source capture–recapture analysis (1CRC). A practical advantage of the 1CRC method is that, compared to the treatment multiplier, it does not require a field sample. Further, estimation based on a single registration has two important advantages: first, it does not require the unverifiable assumption that the two data sources are statistically independent, and second, it does not require the elaborate process of database linkage that is often troubled by privacy regulations, and various technical problems in linking two datasets. A single registration that contains re-admissions circumvents these problems.

In the field of drug research, 1CRC has been applied to estimate the number of dealers and drug users in Quebec (Bouchard & Tremblay, 2005), the size of the cannabis cultivation industry in Quebec (Bouchard, 2007, 2008), and the number of young drug users in Italy (Mascioli & Rossi, 2008). The difference here is that we make use of regression models, whereas these studies include covariates by doing a separate estimation for each subpopulation. Thus, we can take characteristics of drug users, as measured by covariates, into account. An early application of 1CRC was reported

* Corresponding author. Tel.: +31 30 2534688; fax: +31 30 2535797.
E-mail address: P.G.M.vanderHeijden@uu.nl (P.G.M. van der Heijden).

0955-3959/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.drugpo.2013.04.002

Table 1
Frequency distribution of hospital admissions of problem drug users in 2009 in the Netherlands, by gender and four substances of abuse.

Variable	1	2	3	4	5	6	7	8	9	10	11	Total
Gender												
Male	1134	117	28	8	5	1	1	0	0	0	1	1295
Female	346	38	13	3	5	2	1	0	0	1	1	410
Opiates												
Yes	436	75	21	8	2	2	1	0	0	0	2	547
No	1044	80	20	3	8	1	1	0	0	1	0	1158
Cocaine												
Yes	561	62	16	7	7	2	2	0	0	0	1	658
No	919	93	25	4	3	1	0	0	0	1	1	1047
Amphetamines												
Yes	175	11	2	1	1	0	0	0	0	0	0	190
No	1305	144	39	10	9	3	2	0	0	1	2	1515
Cannabis												
Yes	505	40	10	2	1	0	0	0	0	1	0	559
No	975	115	31	9	9	3	2	9	9	9	2	1146

Source: Dutch Hospital Registration (LMR), Dutch Hospital Data Foundation (DHD).

in Smit et al. (1997) for estimating the number of opiate users in the Dutch city of Rotterdam. Since then the method has been developed further (Böhning & van der Heijden, 2009; Cruyff & van der Heijden, 2008; van der Heijden, Bustami, Cruyff, Engbersen, & Van Houwelingen, 2003). We illustrate the method here by showing that for the Netherlands we can produce a national estimate of the number of problem drug users by means of 1CRC by using the Dutch Hospital Registration (LMR), a national registration of hospital admissions held by the Dutch Hospital Data Foundation (DHD).

The methods yield the following results: (i) the hidden number of problem drug users, (ii) the hidden number of users of specific drugs, (iii) a distribution of these hidden numbers over covariates, and (iv) insight into which part of the hidden number is visible in the registration and which part is missed, stratified by the levels of the covariates.

Data source

Each year, the Netherlands National Drug Monitor (NDM) reports the drug situation to the EMCDDA (Van Laar et al., 2012). The NDM relies on the Dutch Hospital Registration (LMR) to monitor drug-related hospital admissions. For each hospital admission, a main diagnosis is registered in the LMR to which one or more secondary diagnoses can be added. Diagnoses linked to drug-related admissions refer to abuse and dependence and are coded to ICD-9. For the registration year 2009, the LMR documented a total of 1705 patients who were hospitalised for a primary or a secondary diagnosis mentioning opiates, cocaine, amphetamines, or cannabis. The distribution of these patients over the four problem drug users groups and the respective frequency distributions are given in Table 1.

Problem users are seen at least once. The aim of the analysis is to estimate the number of problem users with identical covariate values that are seen zero times. Thus, by adding this estimate to the number of observed problem drug users, we find the population size estimate. Given the nature of the LMR, 1CRC will result in estimates of a subgroup of those problem drug users who, due to their problem drug use, are at risk of a clinical hospital treatment.

Methods

We assume that for a single registration it is possible to generate counts for individuals that summarise their capture–recapture histories. In population size estimation models, a key assumption is that a count of an individual is a realisation from a Poisson distribution, which is determined by a single parameter called the Poisson parameter. Obviously, realistic models allow different individuals to have different Poisson parameters, and the models discussed here comply with this. First, however, we discuss the assumption of a Poisson distribution for an individual.

Poisson distribution

An informal definition of a Poisson distribution is as follows. The Poisson distribution is characterised by a Poisson parameter denoted by λ. This parameter λ is used to determine the probability of a given number of events (i.e. the count) under two assumptions: that events occur (i) with an average rate in a fixed interval of time and (ii) independent of the time since the last event. From our analyses we provide two examples of a Poisson distribution in the left panel of Fig. 1: Women aged 45 using opiates, having an estimated Poisson parameter $\lambda = .70$, and women aged 45 using amphetamines, having estimated parameter of $\lambda = .34$. For $\lambda = .70$ (opiates) the probability of being seen at least once is 50. For $\lambda = .34$ (amphetamines), the probability of being seen once is at least 29. It follows that the women using opiates have a larger probability of being seen than the women using amphetamines. Property (i) is less restrictive than it may seem: if the probability of being seen is not constant, the assumption of a Poisson distribution may still be valid (see van der Heijden et al., 2003).

Property (ii) of the Poisson distribution is more restrictive. It means that a change in the individual Poisson parameter is assumed to be unrelated to any prior time seen or not seen. For example, if someone who is using drugs makes a drug-related visit to a hospital, then according to the Poisson assumption this should not increase or decrease his or her probability of making a second drug-related visit. We come back to this point in the discussion.

In 1CRC data, the observed individuals all have a count larger than zero. As an observed count cannot be zero, individuals have so-called 'truncated Poisson distributions'. For the distributions in the left panel of Fig. 1, truncated distributions are displayed in the right panel of Fig. 1.

If there are no covariates, there is only a single Poisson parameter that all individuals have in common (homogeneity assumption). This would then suggest that for a dataset the observed Y-values are drawn from a truncated Poisson distribution and we would have to estimate one single Poisson parameter λ. λ is used to obtain the probability of not being registered, denoted by $P(y = 0|\lambda)$. Using $P(y = 0|\lambda.)$ we would then be able to estimate the part of the population that we did not see. For example, assume that we have seen $n = 250$ individuals with a Poisson parameter for

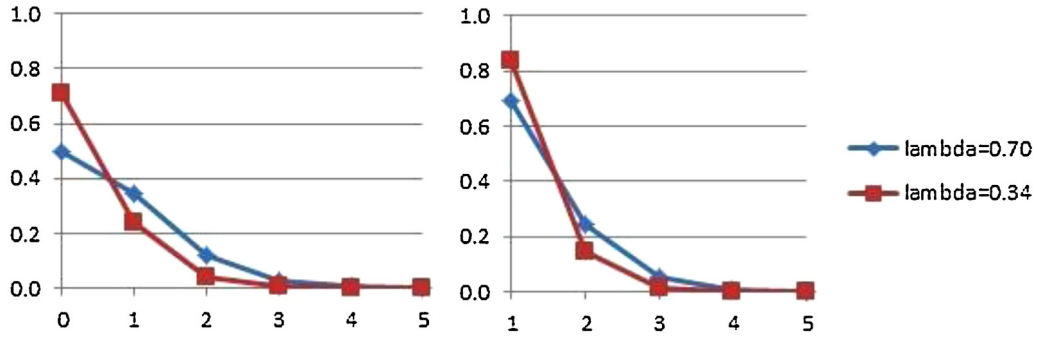

Fig. 1. Poisson distributions for $\lambda = .34$ and $\lambda = .70$. On the horizontal axis the number of recaptures is shown and on the vertical axis the probability. For each line the probabilities add up to 1. The left figure shows untruncated counts (including 0) and the right figure truncated counts, where $P(0)$ is to be estimated.

which $P(y = 0|\lambda.) = .667$. This would mean that $P(y > 0|\lambda.) = .333$, i.e. we observed only one-third of the population, so n refers to one third of the population size. The missed part of the population size is $(.667/.333) * 250 = 500$. The estimated population size would then be equal to the observed part plus the missed part, i.e. $N = 250 + 500 = 750$. In the statistical literature this is known as the Horvitz–Thompson estimator of the population size (see van der Heijden et al., 2003).

Model 1: the truncated Poisson regression model (TPR)

The assumption that all individuals have the same Poisson parameter is unrealistic. In the truncated Poisson regression model this is relaxed into the assumption that the Poisson parameter is a function of covariate values. Thus individual Poisson parameters may differ, but individuals with identical covariates have identical Poisson parameters. Secondly, the Poisson parameter for individual i is related to the covariate values $x_{1i}, x_{2i}, \ldots$ of individual i by a log-linear model, i.e.

$$Log\ \lambda_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots \quad (2)$$

This equation explains the term 'regression' in the name 'truncated Poisson *regression* model'. We can apply the Horvitz–Thompson method on an individual level once the model is estimated and the parameters are known. Using our earlier example, for each of the 250 individuals in the data we have an estimated λ_i-parameter. For each individual i we now calculate $P(y = 0|\lambda_i)/P(y > 0|\lambda_i)$ and this yields the number of missed individuals with covariate values identical to individual i. If we now add up these numbers of missed individuals and add 250, we have the estimated population size derived under the truncated Poisson regression model.

Summarising, we have exchanged the homogeneity assumption for a heterogeneity assumption that allows individuals to be different with regard to their covariates. Incorrectly ignoring heterogeneity leads to an estimated population size that is too low (van der Heijden et al., 2003). Ignoring heterogeneity may occur if important covariates are not used in (2). It is possible to investigate whether heterogeneity is ignored by using a test presented by Gurmu (1991). If this test is significant, then the estimated population size is to be interpreted as a lower-bound for the true population size. Thus the Gurmu test can be used to investigate the fit of the model. An early example of how this model was used in drug research is provided by Smit et al. (1997). An extension to the negative binomial regression model can be found in Cruyff and van der Heijden (2008). This latter model allows for additional heterogeneity of Poisson parameters after the covariates have been taken into account.

Model 2: the Zelterman regression model

Zelterman (1988) proposes not to use the complete distribution of counts to estimate λ but only f_1 and f_2. Thus he derives an estimate of the Poisson parameter

$$\hat{\lambda} = \frac{2f_2}{f_1} \quad (3)$$

and this estimate of λ can be used to estimate $P(0)$, and hence the population size.

Zelterman's proposal has the advantage that it is robust against violations of the Poisson assumption such as unobserved heterogeneity. Also, when interest goes out to the missed count f_0, it makes sense to use the information that is the most close to f_0, i.e. f_1 and f_2, because individuals that make up f_1 and f_2 will be most similar to the individuals that make up f_0. Probably for this reason, and because the estimator is easy to understand, Zelterman's estimator is popular in estimates of drug using populations (see, for example, Van Hest, Grant, Smit, Story, & Richardus, 2007).

Recently Böhning and van der Heijden (2009) generalised (3) so that it can take covariates into account using (2). Due to the robustness property, the Zelterman regression model is a useful competitor of the TPR model when the homogeneity assumption of the Poisson parameters (conditional on the covariates) is violated.

Results

We carried out an analysis using the TPR model and the Zelterman regression model. The covariates used were gender (male = 1, female = 0), age (divided by 10), opiates, cocaine, amphetamines, and cannabis (for each drug separately, yes = 1, no = 0). We note that patients could have been hospitalised because of the use of more than one drug. The dependent variable is the number of times someone was seen in hospital.

The TPR model had a bad fit. The Gurmu (1991) test showed that there was still considerable heterogeneity of Poisson parameters after the six covariates were taken into account (Pearson $X^2 = 28.2$, df is 1, $p < .001$).

For the TPR model the estimate of the population size is 6695 (95% CI is 5779–7611), but given the misfit of the model, the population size estimate is likely to be too small. For completeness we provide the regression parameter estimates in the first two columns of Table 2. The coefficients show that, conditional on the other covariates, the covariates gender, age, opiates and cocaine have a significant contribution, and amphetamines and cannabis do not. These coefficients show the effect on the mean of the natural logarithm of the Poisson parameter, so conditional on the other covariates for males this mean is lower than for females – implying that males have a smaller probability of being seen at all, the mean is higher for older users and for opiate and cocaine users. The

Table 2
Regression parameter estimates for the TPR model (columns 1 and 2) and the Zelterman regression model (ZRM, columns 3 and 4).

	TPR	SE	ZRM	SE
Constant	−2.06	.24***	−3.27	.40***
Male	−.49	.11***	−.05	.20
Age/100	.22	.05***	.12	.08
Opiates	.73	.13***	.90	.24***
Cocaine	.64	.12***	.42	.22
Amphetamines	.01	.23	.01	.36
Cannabis	.24	.16	.28	.26

Source: Dutch Hospital Registration (LMR), Dutch Hospital Data Foundation (DHD).
*** $p < .001$.

Table 3
Observed and estimated population sizes of subgroups and poly-drug use group under the Zelterman regression model.

	Obs.	Estim.	CI	Seen (%)
Male	1295	7991	6299–9682	.16
Female	410	2424	1588–3260	.17
Opiates (yes)	547	1933	1499–2367	.28
Cocaine (yes)	658	3688	2695–4680	.18
Amphetamines (yes)	190	1725	686–2764	.11
Cannabis (yes)	410	4082	2776–5389	.14

Source: Dutch Hospital Registration (LMR), Dutch Hospital Data Foundation (DHD).

opiate users in particular are known to be an ageing population, suffering more and more from chronic disease. This explains why these problem drug users are seen more frequently at the hospital.

The Zelterman regression model uses the observed counts of 1 and 2 to estimate the regression parameters. For the Zelterman regression model, the estimate of the population size is 10,415 (95% CI is 8400–12,429). The parameter estimates are in the last two columns of Table 2. We note that the order of their sizes is identical to the order of their sizes in the TPR model, but that the estimated standard errors under the Zelterman regression model are almost twice as large. As a result only the parameter estimate for opiates is significant, and the parameter estimate for cocaine is borderline significant ($p = .055$).

Table 3 shows how this population size estimate is built up for subgroups. The estimated subpopulation size for males is 7991 and for females it is 2424. The estimated population sizes for opiate, cocaine, amphetamine, and cannabis users are, respectively, 1933, 3688, 1725, and 4082. These four counts do not add up to the estimated population size 10,415, but to 11,428, as there are users who fall into more than one drug category (details about poly-drug use can be obtained upon request from the authors). However, the number of overlapping users is not large.

Which part of the population of drug users at risk of going to hospital is actually seen in hospital? For the full population this is 100 * observed/estimated = 100 * 1705/10,415 = 16%. In the last column of Table 3 percentages are provided for subgroups; such percentages are found by dividing the estimated count for the subgroup by the estimate of the total population 10,415. A much larger part of the population of opiate users compared to the users of other drugs is seen in hospital (namely 28%) and this reflects the significance of the parameter estimate for opiates; for amphetamine users this percentage is smallest (11%).

Discussion and conclusion

'Problem drug use' is a multi-dimensional concept (Conway et al., 2010) that refers to having more or less social, educational and employment, legal and financial problems as well as mental and physical health problems. Here we give an estimate of the subgroup of problem drug users having bad health. Their health problems are so serious that they are in need of an inpatient clinical hospital admission. Some of them do receive the hospital care they need, but others, who are more socially excluded, do not. Some only receive first aid or some medical care from an addiction specialist, others receive no medical care at all.

The total number of problem drug users having bad health was estimated at about 10,415 problem users, of whom about 1933 were estimated to be problem *opiate* users. In a previous study that applied the treatment multiplier, the number of problem opiate users in the Netherlands was estimated at about 17,700 (Cruts & Van Laar, 2010). By combining these figures it can be further estimated that about 11% of the problem opiate users are in such bad health that they require a clinical hospital admission. Moreover, from the problem opiate users who are estimated to be in need of clinical treatment, about 28% are estimated to be seen and registered in the hospital as suffering from drug use or dependence.

We have focused on two population size estimation models that allow for the inclusion of covariates. The presence of covariates is useful for understanding the composition of the population to be estimated, as they allow for the computation of subgroup estimates.

A key assumption in all models is that a change in the individual Poisson parameter is unrelated to any prior time seen or not seen. However, problem drug users are a difficult group of patients, especially in the case of drug-induced psychosis and comorbid personality disorders. For a problem drug user, as well as for the nurses and the doctors, a hospital admission can be a negative experience. Therefore, an initial hospital admission may decrease the chance of a second admission. Moreover, given the current social, financial, and political climate in the Netherlands, problem drug users will experience more and more social exclusion, including less access to health care. If an earlier visit lowers the probability of a second visit, then our estimates will be somewhat high.

Acknowledgments

The Trimbos Institute, Netherlands Institute of Mental Health and Addiction, receives funding from the European Monitoring Centre for Drugs and Drug Addiction (EMCDDA) and from the Netherlands Ministry of Health, Welfare, and Sport (VWS) to monitor the drug situation in the Netherlands. The Dutch Hospital Data Foundation (DHD) delivered the data from the Dutch Hospital Registration (LMR).

Conflict of interest

The authors have no conflict of interest to declare.

References

Böhning, D., & van der Heijden, P. G. M. (2009). A covariate adjustment for zero-truncated approaches to estimating the size of hidden and elusive populations. *The Annals of Applied Statistics*, *3*(2), 595–610.
Bouchard, M. (2007). A capture–recapture model to estimate the size of criminal populations and the risks of detection in a marijuana cultivation industry. *Journal of Quantitative Criminology*, *23*(3), 221–241.
Bouchard, M. (2008). Towards a realistic method to estimate cannabis production in industrialized countries. *Contemporary Drug Problems*, *35*(2B), 291–320.
Bouchard, M., & Tremblay, P. (2005). Risks of arrest across drug markets: A capture–recapture analysis of "hidden" dealer and user populations. *Journal of Drug Issues*, *35*(4), 733–754.
Conway, K. P., Levy, J., Vanyukov, M., Chandler, R., Rutter, J., Swan, G. E., & Neale, M. (2010). Measuring addiction propensity and severity: The need for a new instrument. *Drug and Alcohol Dependence*, *111*, 4–12.
Cruts, A. A. N, & Van Laar, M. W. (2010). *Aantal problematische harddrugsgebruikers in Nederland [Number of problem harddrug users in the Netherlands]*. Utrecht, the Netherlands: Trimbos Institute, Netherlands Institute of Mental Health and Addiction.
Cruyff, M. J. L. F., & van der Heijden, P. G. M. (2008). Point and interval estimation of the population size using a zero-truncated negative binomial regression model. *Biometrical Journal*, *50*(6), 1035–1050.

EMCDDA. (2011). *Annual report 2011: The state of the drugs problem in Europe*. Luxembourg: Publications Office of the European Union.

Gurmu, S. (1991). Tests for detecting overdispersion in the positive Poisson regression model. *Journal of Business and Economic Statistics*, *9*, 215–222.

Mascioli, F., & Rossi, C. (2008). Capture–recapture methods to estimate prevalence indicators for the evaluation of drug policies. *Bulletin on Narcotics*, *LX*(1/2), 5–25.

Smit, F., Toet, J., & van der Heijden, P. (1997). Estimating the number of opiate users in Rotterdam using statistical models for incomplete count data. In G. Hay, N. McKeganey, & E. Birks (Eds.), *Final report EMCDDA project methodological pilot study of local level prevalence estimates*. Glasgow: Centre for Drug Misuse Research, University of Glasgow/EMCDDA.

Temürhan, M., Meijer, R., Choenni, S., Van Ooyen-Houben, M., Cruts, G., & Van Laar, M. Capture–recapture method for estimating the number of problem drug users: the case of the Netherlands, submitted for publication.

van der Heijden, P. G. M., Bustami, R., Cruyff, M. J. L. F., Engbersen, G., & Van Houwelingen, H. C. (2003). Point and interval estimation of the truncated Poisson regression model. *Statistical Modelling*, *3*, 305–322.

Van Hest, N. H. A., Grant, A. D., Smit, F., Story, A., & Richardus, J. H. (2007). Estimating infectious diseases incidence: Validity of capture–recapture analysis and truncated models for incomplete count data. *Epidemiology and Infection*, *136*, 14–22.

Van Laar, M. W., Cruts, A. A. N., Van Ooyen-Houben, M. M. J., Van Gageldonk, A., Croes, E. A., Meijer, R. F., & Ketelaars, A. P. M. (2012). *The Netherlands Drug Situation 2011: Report to the EMCDDA by the Reitox National Focal Point.* Utrecht, The Netherlands: Trimbos Institute, Netherlands Institute of Mental Health and Addiction.

Zelterman, D. (1988). Robust estimation in truncated discrete distributions with application to capture–recapture experiments. *Journal of Statistical Planning and Inference*, *18*, 225–237.